Research on an efficient Tracking and Regulation Method for Hot Issues in Network Forums

Li-Jie Cui[1], Wei Liu[2], Guang-Yi Tang[1]

[1] School of Software and Engineering , Harbin University of Science and Technology, Harbin, China
[2]Harbin Institute of Technology Software Engineering Co. Ltd, Harbin, China
Li-Jie Cui: andyclj1977@163.com

Abstract. Hot issues are problems that trigger the concern and interest of people. To control public opinions, we should track, search, and regulate hot issues to safeguard social fairness and justice. Thus, analyzing and regulating online forums could strengthen the management and monitoring of public opinion. An efficient tracking and regulation method for hot posts is needed to control public opinions. This paper presents a fast and effective method that can discover hot posts and a mathematical method that can analyze the evolutionary trend of these posts. This paper is designed to lay a foundation and pave the way for the next regulation system. On the basis of the analysis results, the proposed method is rapid, feasible, and can obtain ideal experimental results.

Keywords: network forum; hot issue; public opinion; information dissemination

1 Introduction

The core elements of information dissemination and public opinions in the Internet include hot issues, focus, sensitive points, frequency, etc. Here, “hot issues” refer to topics subjected to moral judgment. [1]Hot issues are problems that trigger the concern and interest of people. These issues even affect social stability and may impede the construction of a harmonious society.[2] Thus, to control public opinions, we should track, search, and regulate hot issues to safeguard social fairness and justice.

2 Principle of related methods

2.1 Data collection

We have mined a total of 18 753 posts from February 2010 to December 2010 in the "People Voice" section of the Tianya Club. Each post includes the following information: post title, subordinated section, number of visits, number of replies, author, publication date, post contents, and comment replies. Among these data, we focus on the number of visits and number of replies of posts

2.2 Data Analysis

We have extracted the number of visits and number of replies for each post whose values constituted a tuple (number of visits, number of replies). We can obtain information regarding the *i*th post by checking the tuple. The extracted attributes, namely, number of visits and number of replies, are projected onto the following Cartesian coordinate system (Figure 1):

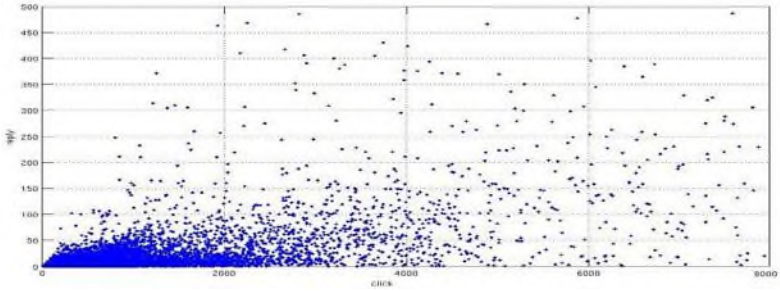

Fig. 1. Distribution of number of visits versus number of replies

In Figure 1, the *x*-axis represents the number of visits, whereas the *y*-axis represents the number of replies. In general, the distribution of number of visits and number of replies adhere to the following patterns:

1. Few visits. These types of posts are called "cold posts," which are regarded as meaningless posts.

2. Large number of visits, but few replies. This case corresponds to the lower right quadrant of Figure1. In this situation, we cannot evaluate whether a post is a hot post by using only the number of visits.

3. Large number of visits and replies. These posts are located in the upper right quadrant of Figure 1. These posts are the "hot posts" needed in this study.

2.3 Methods and results

To extract hot posts, we can score all the posts and then select the top *N* posts according to need. The following scoring formula is used in scoring these posts:

$$S(p_i) = w_1 \frac{x_i}{average(x)} + w_2 \frac{y_i}{average(y)} + w_3 \frac{y_i}{x_i} \Big/ \max(a) \quad (1)$$

where $S(p_i)$ represents the score of the *i*th post, $average(x)$ represents the average number of visits, $average(y)$ represents the average number of replies, $\max(a)$ represents the maximum ratio of the number of visits and number of replies in

all the tuples, and w_1, w_2, w_3 represent the weighting factors. Considering that the number of replies can better reflect the probability whether a post may result in a debate, we assume that $w_1 < w_2$. We can consider a third factor if the scores of the posts cannot be determined by the number of visits and number of replies. By adjusting the proportion of the number of replies, we can obtain $0 < \frac{y_i x_i}{\max(a)} < 1$.

3 Development trend analysis

3.1 Preprocessed data

This thread lasted for 58 days. Groove marks are created for each day by using an array to segregate comments according to publication date. The subscript of the array is mapped to the *x*-axis, and the value of the array becomes the value of the *y*-axis. Figure 2 shows the discrete points plotted in the Cartesian coordinate system.

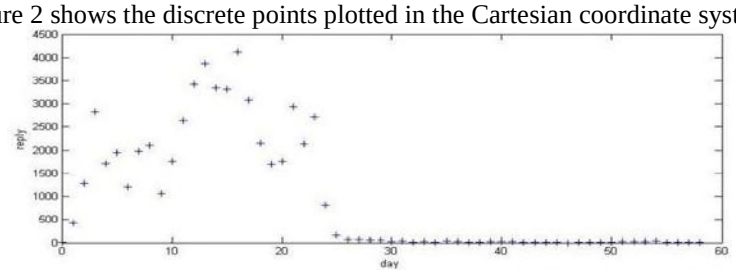

Fig. 2. Number of replies per day over time.

In Figure 2, the *x*-axis represents the days that passed since the creation of the post, whereas the *y*-axis represents the number of corresponding responses. The popularity of this post tends to decrease and level off, and soon the post "sinks." After removing noise points, we can utilize a suitable curve that fits these discrete points. The result is presented below (Figure 3).

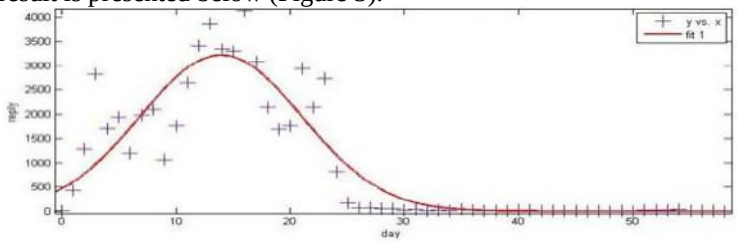

Fig. 3. Number of replies per day distributed over a Gaussian curve.

3.2 Selection and evaluation of mathematical models

We can use the Gaussian curve as a tool for fitting the data curve of other posts. The Gaussian model is expressed as follows:

$$y(x) = a \cdot e^{-\left(\frac{x-b}{c}\right)^2} \quad (2)$$

The following parameter values are obtained:

$$a = 3218(2849,3586)$$
$$b = 13.92(12.99,14.85)$$
$$c = 9.984(8.167,11.35) \quad (3)$$

The above values have a confidence interval of 0.95.We can obtain the following:

$$y(x) = 3218 \cdot e^{-\left(\frac{x-13.92}{9.984}\right)^2} \quad (4)$$

The evaluation of the fitting effect is detailed below:

1. $SSE = 1.548e+07$. ^{SSE} , which is the sum of squares of an error term, reflects the discrete status of the observed values of each sample.
2. $R\text{-}square = 0.8339$. $R\text{-}square$ is the fitting coefficient.
3. $RMSE = 525.7$. ^{RMSE} is the root mean square error, which is a numerical index that can be used to measure accuracy.

3.3 Core analysis method

Let $y'(x) = 0$; then, we can obtain the largest extreme point and set it as ^{xm} . The following two conditions are necessary in analyzing posts:

1. $y = f(x)$ has no extreme points;
2. If $\$ x > x_m$, then $y'(x) > 0$.

If condition 1 is established, then $y'(x) > 0$ or $y'(x) < 0$ and $^{y(x)}$ is monotonic. If condition 2 is satisfied, then $y'(x) > 0$, and $^{y(x)}$ is monotonically increasing. When both conditions are satisfied, the heating degree of the post has an upward trend, and we should continue to focus on the trend and evolution of this post.

If point ^{xn} exists such that $y'(x_n) = 0$ holds true, then ^{xn} is a turning point in the heating trend. We then determine the largest $^{x\max}$ that can establish $y'(x_n) = 0$. By studying ^{xt} , which satisfies $x_t > x_{\max}$, we can denote that if $y'(x_t) < 0$, then the heat of the post experiences a downward trend.

Correspondingly, if $y'(x_t) > 0$, then the heat degree has an upward trend, and we should be concerned on the future trend of this post.

4 Conclusion

This paper presents a fast and effective method that can discover hot posts and a mathematical method that can analyze the evolutionary trend of these posts. This paper is designed to lay a foundation and pave the way for the next regulation system.

References

1. Hai-Guang Xie ,Zhong-Ren Chen .Internet content and public opinion depth analysis mode.China Youth Journal for Political Sciences.3.2006
2. Bin Lu .Automatic discovery and analysis of the Internet public opinion hotspot Research and Implementation.Master's degree thesis of Department of Computer Application Technology of Beijing University